

# Tuning Lasso for sup-norm optimality

Michaël Chichignoud

Johannes Lederer

Seminar for Statistics

Department of Statistical Science

ETH Zürich

Cornell University

CH-8092 Zürich

Ithaca, NY 14853

michael.chichignoud@gmail.com

johannesleder@cornell.edu

Martin Wainwright

Department of Statistics and Department of Electrical Engineering and Computer Sciences

University of California at Berkeley

Berkeley, CA 94720

wainwrig@stat.berkeley.edu

October 2, 2014

## Abstract

We introduce novel schemes for tuning parameter calibration in high-dimensional linear regression with Lasso. These calibration schemes are inspired by Lepski's method for bandwidth adaptation in non-parametric regression and are the first calibration schemes that are equipped with both theoretical guarantees and fast algorithms. In particular, we develop optimal finite sample guarantees for sup-norm performance and give algorithms that consist of simple tests along a single Lasso path. Moreover, we show that false positives can be safely reduced without increasing the number of false negatives. Applying Lasso to synthetic data and to real data, we finally demonstrate that the novel schemes can rival standard schemes such as Cross-Validation in speed as well as in sup-norm and variable selection performance.

KEYWORDS: Lasso, tuning parameter, high-dimensional regression

## 1. INTRODUCTION

Regularized estimators such as Lasso (Tibshirani 1996), Square-Root Lasso (Belloni, Chernozhukov and Wang 2011; Städler, Bühlmann and van de Geer 2010; Sun and Zhang 2012), MCP (Zhang 2010), and SCAD (Fan and Li 2001) hinge on tuning parameters. However, standard calibration schemes for these tuning parameters are computationally infeasible or intractable for finite sample theory. As a consequence, the calibration of regularized estimators is a crucial, unsolved problem.

We focus on the calibration of Lasso for linear regression, where the tuning parameter needs to suit both the noise distribution and the design matrix (van de Geer and Lederer 2013; Hebiri and Lederer 2013; Dalalyan, Hebiri and Lederer 2014). Calibration schemes for this setting are typically based on Cross-Validation or BIC-type criteria. However, Cross-Validation requires extensive computations and lacks for finite sample guarantees. BIC-type criteria, on the other hand, are computationally simpler but also lack for finite sample guarantees. Another approach is to replace Lasso with Square-Root Lasso or TREX (Lederer and Müller 2014); however, Square-Root Lasso still contains a tuning parameter that needs to be calibrated to certain aspects of the model, and the theory for TREX is currently fragmentary. Therefore, new insights into the calibration of Lasso are necessary.

In this paper, we introduce novel calibration schemes for minimal sup-norm loss of Lasso: Adaptive Validation for  $\ell_\infty$  ( $AV_\infty$ ) and Adaptive Validation With Multiple Constants ( $AV_\infty^m$ ).  $AV_\infty$  and  $AV_\infty^m$  consist of tests that are inspired by Lepski’s method for nonparametric regression (Lepski 1990; Lepski, Mammen and Spokoiny 1997) and provide, in strong contrast to standard calibration schemes, both **optimal theoretical guarantees** and **fast computations**. In Section 2, we introduce  $AV_\infty$  and  $AV_\infty^m$  and present the following:

- Optimal finite samples guarantees for the calibration of Lasso with respect to sup-norm loss in a broad class of linear regression models (Theorem 1 and Theorem 2);
- A safe thresholding procedure to reduce false positives (Remark 2);
- A simple and fast algorithm (Algorithm 1).

These results show that  $AV_\infty$  and  $AV_\infty^m$  can maximize the Lasso performance for sup-norm loss and variable selection. In Section 3, we then support these results with applications to synthetic data and to biological data, and in Section 4, we finally conclude with a discussion of our approach.

## 2. METHODOLOGY

### 2.1 Framework

We study tuning parameter calibration for Lasso for linear regression models that can contain many predictors and correlated, heavy-tailed noise. More specifically, we assume that the data  $(Y, X)$  with outcome  $Y \in \mathbb{R}^n$  and design matrix  $X \in \mathbb{R}^{n \times p}$  is distributed according to a linear regression model

$$Y = X\beta^* + \varepsilon, \tag{Model}$$

where  $\beta^* \in \mathbb{R}^p$  is the regression vector and  $\varepsilon \in \mathbb{R}^n$  random noise. We allow for  $p$  larger than  $n$  and impose only the second moment assumption  $\sup_{i \in \{1, \dots, n\}} \mathbb{E}[\varepsilon_i^2] < \infty$  on the noise. Lasso for the above model is the family of estimators (see (Bühlmann and van de Geer 2011) for an overview)

$$\hat{\beta}_\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right\}, \tag{Lasso}$$

indexed by the tuning parameter  $\lambda > 0$  that determines the amount of regularization.

To ease the exposition in the sequel, we also set some conventions and notations: The design matrix can be fixed or random; for the latter case, the results are to be understood conditionally on  $X$ . The indicator of events is denoted by  $\mathbb{1}\{\cdot\} \in \{0, 1\}$ , the cardinality of a set by  $|\cdot|$ , the sup-norm (maximum norm) of  $\mathbb{R}^p$ -vectors by  $\|\cdot\|_\infty$ , the number of nonzero entries by  $\|\cdot\|_0$ , and the  $\ell_1$ - and  $\ell_2$ -norms by  $\|\cdot\|_1$  and  $\|\cdot\|_2$ , respectively. The support of  $\beta^*$  is finally denoted by  $S$ , and, for any vector  $\beta \in \mathbb{R}^p$ , the components in  $S$  and in the complement of  $S$  are denoted by  $\beta_S \in \mathbb{R}^{|S|}$  and  $\beta_{S^c} \in \mathbb{R}^{|S^c|}$ , respectively.

### 2.2 Notion of optimality

Having specified the framework, we can now introduce the notion of optimality for sup-norm loss. We stress first that tuning parameter calibration needs to be tailored to the loss: For example, a tuning parameter optimal for prediction loss  $\|X\hat{\beta}_\lambda - X\beta^*\|_2$  can be completely unsuitable for sup-norm loss  $\|\hat{\beta}_\lambda - \beta^*\|_\infty$  (and vice versa). For fixed tuning parameters, sup-norm bounds for Lasso have been developed in (Bunea 2008; Lounici 2008). Combining results in (Lounici 2008), we obtain the following bound:

**Lemma 1** (See (Lounici 2008)). Assume that the Gram matrix  $G := X^\top X/n$  is normalized such that  $G_{ii} = 1$  for all  $i \in \{1, \dots, p\}$ , and assume that  $G$  fulfills the mutual coherence condition

$$\max_{i \neq j} |G_{ij}| \leq \frac{1}{7c\|\beta^*\|_0} \quad (\text{Mutual coherence})$$

for a constant  $c > 1$ . Then, on  $\mathcal{T}_\lambda := \{4\|X^\top \varepsilon\|_\infty/n \leq \lambda\}$ , the following sup-norm bound for Lasso holds:

$$\|\hat{\beta}_\lambda - \beta^*\|_\infty \leq \tilde{a}_c \lambda,$$

where  $\tilde{a}_c := \frac{3}{4} \left(1 + \frac{16}{7(c-1)}\right)$ .

This provides an (essentially sharp) upper bound for the sup-norm performance of Lasso with a sufficiently large tuning parameter. If the noise  $\varepsilon$  was accessible, one would be able to select tuning parameters that are optimal for this bound:

**Definition 1** (Oracle tuning parameter). For any constant  $b > 0$ , we define the (inaccessible) oracle tuning parameter as

$$\lambda_b^* := \min_{\lambda > 0} \{\mathbb{P}(\mathcal{T}_\lambda) \geq 1 - b\}.$$

**Corollary 1** (Optimal guarantee). Assume that the assumptions in Lemma 1 are met. Then, for any constant  $b > 0$ , the following sup-norm guarantee for Lasso with the oracle tuning parameter  $\lambda_b^*$  holds with probability at least  $1 - b$ :

$$\|\hat{\beta}_{\lambda_b^*} - \beta^*\|_\infty \leq \tilde{a}_c \lambda_b^*.$$

Corollary 1 states the best guarantees that can be obtained (in the sense of Lemma 1) by any calibration scheme. We therefore call a calibration scheme optimal if it provides - up to constants - the guarantees in Corollary 1.

### 2.3 Adaptive calibration

To obtain optimal calibration in the above sense, we now introduce Adaptive Calibration for  $\ell_\infty$  ( $AV_\infty$ ):

**Definition 2** ( $AV_\infty$ ). For a fixed but arbitrary  $a_c \geq \tilde{a}_c$ , Adaptive Calibration for  $\ell_\infty$  ( $AV_\infty$ ) selects the data-driven tuning parameter

$$\hat{\lambda} := \min \left\{ \lambda > 0 : \sup_{\lambda', \lambda'' \geq \lambda} \left[ \frac{\|\hat{\beta}_{\lambda'} - \hat{\beta}_{\lambda''}\|_\infty}{\lambda' + \lambda''} - a_c \right] \leq 0 \right\}.$$

Because  $\hat{\beta}_\lambda = 0$  for  $\lambda$  sufficiently large,  $AV_\infty$  is always well defined. The definition is based on tests for sup-norm differences of Lasso estimates with different tuning parameters. These tests are inspired by Lepski's method (Lepski 1990; Lepski et al. 1997) for bandwidth selection in nonparametric, pointwise regression, see also (Chichignoud and Lederer 2014). We stress that Definition 2 does not require prior knowledge about the regression vector or the noise; in particular, Definition 2 does not require knowledge about the noise variance. We also note that the constant  $a_c$  relates to the constant  $\tilde{a}_c$  in Lemma 1 and, therefore, is specified by the theoretical bounds at hand (in particular,  $a_c$  is *not* a tuning parameter), see Section 3 for details. We finally mention that in the definition and all following results, the tuning parameters can be restricted to subsets of  $(0, \infty)$  (for example, to a grid).

We now provide a simple and fast algorithm for  $AV_\infty$ . For this, let  $N \in \mathbb{N}$  be an integer,  $\lambda_{\max} := 2\|X^\top Y\|_\infty/n$  the smallest tuning parameter for which  $\hat{\beta}_\lambda$  equals zero, and  $0 < \lambda_1 < \dots < \lambda_N = \lambda_{\max}$  a tuning parameter grid. The tests in Definition 2 then correspond to the binary random variables

$$\hat{t}_{\lambda_j} := \prod_{k=j, \dots, N} \mathbb{1} \left\{ \frac{\|\hat{\beta}_{\lambda_j} - \hat{\beta}_{\lambda_k}\|_\infty}{\lambda_j + \lambda_k} - a_c \leq 0 \right\}$$

for  $j \in \{1, \dots, N\}$ . The  $AV_\infty$  tuning parameter  $\hat{\lambda}$  on the tuning parameter grid can now be computed as follows:

**Data:**  $\hat{\beta}_{\lambda_1}, \dots, \hat{\beta}_{\lambda_N}, a_c$

**Result:**  $\hat{\lambda}^{\text{grid}} \in (0, \lambda_{\max}]$

Set initial index:  $j \leftarrow N$

**while**  $(\hat{t}_{\lambda_{j-1}} \neq 0)$  *and*  $(j > 1)$  **do**

  | Update index:  $j \leftarrow j - 1$

**end**

Set output:  $\hat{\lambda}^{\text{grid}} \leftarrow \lambda_j$

**Algorithm 1:** Algorithm for  $AV_\infty$  in Definition 2.

This algorithm can be readily implemented and only requires the computation of one Lasso solution path. In strong contrast,  $k$ -fold Cross-Validation requires the computation of  $k$  solution paths. This implies that Lasso with  $AV_\infty$  can be computed about  $k$  times faster than Lasso with  $k$ -fold Cross-Validation.

## 2.4 Optimality of $AV_\infty$

The following result now shows that  $AV_\infty$  introduced above provides optimal calibration for sup-norm loss:

**Theorem 1** (Optimality of  $AV_\infty$ ). *Assume that the assumptions in Lemma 1 are met. Then, for any constant  $b > 0$ , the following bounds for Lasso with  $AV_\infty$  hold with probability at least  $1 - b$ :*

$$\|\hat{\beta}_{\hat{\lambda}} - \beta^*\|_\infty \leq 3a_c\lambda_b^* \quad \text{and} \quad \hat{\lambda} \leq \lambda_b^*.$$

**Remark 1** (Optimality). *A comparison of the sup-norm guarantee in Theorem 1 with the guarantee in Corollary 1 shows that  $AV_\infty$  is optimal (up to a factor 3). For standard calibration schemes (including Cross-Validation, for example), no comparable guarantees are available in the literature. In fact, we are not aware of **any** finite sample guarantees for standard calibration schemes.*

**Remark 2** (Thresholding). *The bound for the  $AV_\infty$  tuning parameter  $\hat{\lambda}$  can be exploited to safely remove false positives: Thresholding  $\hat{\beta}_{\hat{\lambda}}$  by  $3a_c\hat{\lambda}$  can reduce the number of false positives without decreasing the number of true positives, given that the smallest non-zero entries of the regression vector  $\beta^*$  are (in absolute value) larger than  $3a_c\lambda_b^*$ . In contrast, bounds for tuning parameters provided by standard calibration schemes are not available in the literature, such that thresholding based on these tuning parameters might decrease the number of true positives.*

**Remark 3** (Scope). *We first point out that Theorem 1 provides - in contrast to asymptotic results or results with unspecified constants - explicit guarantees for arbitrary sample sizes. Moreover, Theorem 1 does not presume prior knowledge about the noise distribution or the regression vector and allows for correlated, heavy-tailed noise. Finally, as we show in Section 2.5, the mutual coherence condition on the design matrix  $X$  can be considerably relaxed.*

*On the other hand, our approach requires a bound of the form as in Lemma 1; different bounds that contain, for example, an additional factor  $\|\beta^*\|_0$  or  $\|\beta^*\|_1$  cannot be used. Therefore, our approach cannot be directly transferred to calibration for prediction loss or  $\ell_2$ -estimation loss.*

We finally present a proof of Theorem 1.

*Proof of Theorem 1.* We will show that on the event  $\mathcal{T}^* := \{4\|X^\top \varepsilon\|_\infty/n \leq \lambda_b^*\}$ , the desired inequalities hold. From this, the claim follows noting that  $\mathbb{P}(\mathcal{T}^*) \geq 1 - b$  by the definition of the

oracle tuning parameter  $\lambda_b^*$ .

*Claim 1:* On  $\mathcal{T}^*$ , it holds that  $\hat{\lambda} \leq \lambda_b^*$ .

We proof Claim 1 by contradiction. To this end, we assume that  $\hat{\lambda} > \lambda_b^*$ . Then, the definition of the data-driven tuning parameter  $\hat{\lambda}$  implies that there are two tuning parameters  $\lambda', \lambda'' \geq \lambda_b^*$  such that

$$\|\hat{\beta}_{\lambda'} - \hat{\beta}_{\lambda''}\|_\infty > a_c(\lambda' + \lambda'').$$

On the other hand, the inclusion  $\mathcal{T}^* \subset \mathcal{T}_{\lambda'}, \mathcal{T}_{\lambda''}$  holds. We can therefore apply Lemma 1 with  $\lambda', \lambda''$  and the triangle inequality to deduce that on  $\mathcal{T}^*$

$$\|\hat{\beta}_{\lambda'} - \hat{\beta}_{\lambda''}\|_\infty \leq \|\hat{\beta}_{\lambda'} - \beta^*\|_\infty + \|\beta^* - \hat{\beta}_{\lambda''}\|_\infty \leq a_c(\lambda' + \lambda''),$$

which contradicts the previous display and therefore concludes the proof of Claim 1.

*Claim 2:* On  $\mathcal{T}^*$ , it holds that  $\|\hat{\beta}_{\hat{\lambda}} - \beta^*\|_\infty \leq 3a_c\lambda_b^*$ .

We use Claim 1 and the definition of the data-driven tuning parameter to prove Claim 2. We first note that the definition of the data-driven tuning parameter  $\hat{\lambda}$  implies (since on  $\mathcal{T}^*$ ,  $\hat{\lambda} \leq \lambda_b^*$  according to Claim 1)

$$\|\hat{\beta}_{\hat{\lambda}} - \hat{\beta}_{\lambda_b^*}\|_\infty \leq a_c(\hat{\lambda} + \lambda_b^*).$$

We can then apply Lemma 1 and the triangle inequality to deduce

$$\|\hat{\beta}_{\hat{\lambda}} - \beta^*\|_\infty \leq \|\hat{\beta}_{\hat{\lambda}} - \hat{\beta}_{\lambda_b^*}\|_\infty + \|\hat{\beta}_{\lambda_b^*} - \beta^*\|_\infty \leq a_c(\hat{\lambda} + \lambda_b^*) + a_c\lambda_b^* \leq 3a_c\lambda_b^*.$$

This completes the proof of Claim 2 and, therefore, concludes the proof of Theorem 1.  $\square$

## 2.5 Beyond mutual coherence

In the following, we show that a modification of  $\text{AV}_\infty$  can permit us to assume considerably weaker assumptions on the design matrix. The strict assumptions on the design matrix for  $\text{AV}_\infty$  root in the limitations of the Lasso sup-norm bound in Lemma 1. We therefore derive a different sup-norm bound with weaker assumptions. For this, we fix for each index  $j \in \{1, \dots, p\}$  a deterministic vector

$$\eta^j \in \arg \min_{\substack{\beta \in \mathbb{R}^p \\ \beta_j = -1}} \left\{ \frac{\|X\beta\|_2^2}{n} + u_j \|\beta\|_1 \right\}$$

invoking Lasso with a given tuning parameter  $u_j > 0$ . (Note that this definition corresponds to a regression of the  $j$ th column of the design matrix on the set of all other columns.) To simplify the notation, we then define the norms  $\|a\|_j := |a_j|$  and  $\|a\|_{-j} := \sum_{i \neq j} |a_i|$  for any vector  $a$ . We now state the recent result (van de Geer, S. 2014, Lemma 2.1):

**Lemma 2** (See (van de Geer, S. 2014)). *For all indices  $j \in \{1, \dots, p\}$  and all tuning parameters  $\lambda, u_j > 0$ , it holds that*

$$\|\hat{\beta}_\lambda - \beta^*\|_j \leq \tilde{d}_j \left( \frac{\|X^\top \varepsilon\|_\infty}{n} + \frac{u_j \|\hat{\beta}_\lambda - \beta^*\|_{-j}}{2\|\eta^j\|_1} + \frac{\lambda}{2} \right),$$

where for each  $j \in \{1, \dots, p\}$

$$\tilde{d}_j := \frac{\|\eta^j\|_1}{\|X\eta^j\|_2^2/n + u_j\|\eta^j\|_{-j}/2}.$$

This result provides a specific bound for each coordinate of Lasso. Together with (Bühlmann and van de Geer 2011, Theorem 6.1), these bounds lead to the following result (the proof is straightforward and therefore omitted):

**Lemma 3** (Lasso bound with multiple constants). *Assume that the Gram matrix  $G := X^\top X/n$  is normalized such that  $G_{ii} = 1$  for all  $i \in \{1, \dots, p\}$ , and assume that  $G$  fulfills the compatibility condition*

$$t \geq \min_{\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1} \left\{ \frac{\sqrt{|S|}\|X\beta\|_2}{\sqrt{n}\|\beta_S\|_1} \right\} \quad (\text{Compatibility})$$

for a constant  $t > 0$ . Additionally, assume that

$$\sup_{j \in \{1, \dots, p\}} \frac{u_j |S|}{t^2 \|\eta^j\|_1} \leq \frac{1}{\log(n)}.$$

Then, on  $\mathcal{T}_\lambda := \{4\|X^\top \varepsilon\|_\infty/n \leq \lambda\}$ , the following bound for Lasso and for any index  $j \in \{1, \dots, p\}$  holds:

$$\|\hat{\beta}_\lambda - \beta^*\|_j \leq \left( \frac{3}{4} + \frac{1}{\log(n)} \right) \tilde{d}_j \lambda.$$

We are now ready to introduce a corresponding calibration scheme and state its optimality:

**Definition 3** ( $AV_\infty^m$ ). *For fixed but arbitrary  $d_1 \geq \tilde{d}_1, \dots, d_p \geq \tilde{d}_p$ , Adaptive Validation With Multiple Constants ( $AV_\infty^m$ ) selects the data-driven tuning parameter*

$$\tilde{\lambda} := \min \left\{ \lambda > 0 : \sup_{\lambda', \lambda'' \geq \lambda} \max_{j \in \{1, \dots, p\}} \left[ \frac{\|\hat{\beta}_{\lambda'} - \hat{\beta}_{\lambda''}\|_j}{\lambda' + \lambda''} - d_j \left( \frac{3}{4} + \frac{1}{\log(n)} \right) \right] \leq 0 \right\}.$$



**Theorem 2** (Optimality of  $AV_\infty^m$ ). *Assume that the assumptions in Lemma 3 are met. Then, for any constant  $b > 0$ , the following bounds for Lasso  $AV_\infty^m$  hold with probability at least  $1 - b$  for all indices  $j \in \{1, \dots, p\}$ :*

$$\tilde{\lambda} \leq \lambda_b^* \quad \text{and} \quad \|\hat{\beta}_{\tilde{\lambda}} - \beta^*\|_j \leq 3 \left( \frac{3}{4} + \frac{1}{\log(n)} \right) d_j \lambda_b^*.$$

**Remark 4** (Comparison with  $AV_\infty$ ). *For orthogonal designs, where  $n = p$  and  $X^\top X/n = \mathbf{I}_p$ , the bounds for  $AV_\infty$  and  $AV_\infty^m$  are essentially equal. Indeed, choosing  $u_j = 0$  (in Lemma 3) and  $c = \infty$  (in Lemma 1) yields for all  $\lambda \geq 4\|X^\top \varepsilon\|_\infty/n$  the bounds*

$$\|\hat{\beta}_\lambda - \beta^*\|_\infty \leq \left( \frac{3}{4} + \frac{1}{\log(n)} \right) \lambda \quad \text{and} \quad \|\hat{\beta}_\lambda - \beta^*\|_\infty \leq \frac{3}{4} \lambda,$$

*respectively.*

*For more general designs, however, we choose  $u_j = \mathcal{O}(\sqrt{\log(p)/n})$  (van de Geer, S. 2014) and recall the compatibility condition in Lemma 3 is considerably weaker than the mutual coherence condition in Lemma 1, see (van de Geer and Bühlmann 2009).*

*Moreover, the bounds for  $AV_\infty^m$  are, in contrast to the bounds for  $AV_\infty$ , tailored to each  $j \in \{1, \dots, p\}$ . As shown in Section 3, this tailoring can increase the estimation accuracy.*

### 3. SIMULATIONS

#### 3.1 Synthetic data

We consider  $n = 100$  observations and  $p \in \{200, 500, 2000\}$  parameters. We then sample each row of the design matrix  $X \in \mathbb{R}^{n \times p}$  from the  $p$ -dimensional normal distribution with mean 0 and covariance matrix  $(1 - u)\mathbf{I} + u\mathbf{1}\mathbf{1}^\top$ , where  $\mathbf{I}$  is the identity matrix,  $\mathbf{1} := (1, \dots, 1)(1, \dots, 1)^\top$  is the matrix of ones, and  $u \in \{0, 0.3\}$  is the magnitude of the mutual correlations. For the entries of the noise  $\varepsilon \in \mathbb{R}^n$ , we take the one-dimensional normal distribution with mean 0 and variance 1. The entries of  $\beta^*$  are set to 0 except for 8 uniformly random chosen entries that are each set to 1 or  $-1$  with equal probability. The entire vector  $\beta^*$  is then rescaled to give a signal to noise ratio  $\beta^{*\top} X^\top X \beta^*/n$  equal to 15. The grid of tuning parameters finally consists of 50 points  $\{\lambda_{\max}/50, 2\lambda_{\max}/50, \dots, \lambda_{\max}\}$  with  $\lambda_{\max} := 2\|X^\top Y\|_\infty/n$ . We finally conduct 500 experiments for each set of parameters and report the corresponding means and standard deviations (the latter is stated in brackets).

We compare the sup-norm and variable selection performance of the following four procedures:

- Best: Lasso with the best (but in practice unknown) tuning parameter for  $\ell_\infty$  loss;
- Cross-Validation: Lasso with 10-fold Cross-Validation;
- $AV_\infty$ : Lasso with  $AV_\infty$  and  $a_c = 0.75$ ;
- $AV_\infty^m$ : Lasso with  $AV_\infty^m$  and  $d_1, \dots, d_p$  obtained via a prior Lasso with  $u_j = \sqrt{\log(p)/n}$ .

Note that the constants  $a_c, d_1, \dots, d_p$  are set to values suggested by the theory (cf. Lemmas 1 and 3).

**Sup-norm loss:** In Tables 1 and 2, we compare the  $\ell_\infty$  loss of the four procedures. We observe that both  $AV_\infty$  and  $AV_\infty^m$  rival Cross-Validation in all simulations. Moreover, the simulations indicate that  $AV_\infty^m$  can slightly outperform  $AV_\infty$  in correlated settings (this becomes more pronounced when  $u$  is increased further). Finally, we mention that the same conclusions can be drawn for other, possibly heavy tailed noise distributions (data not shown).

	Best	Cross-Validation	$AV_\infty$	$AV_\infty^m$
$p = 200, u = 0$	0.31 ( $\pm 0.07$ )	0.35 ( $\pm 0.08$ )	0.34 ( $\pm 0.07$ )	0.34 ( $\pm 0.08$ )
$p = 500, u = 0$	0.38 ( $\pm 0.09$ )	0.42 ( $\pm 0.09$ )	0.40 ( $\pm 0.10$ )	0.40 ( $\pm 0.10$ )
$p = 2000, u = 0$	0.52 ( $\pm 0.13$ )	0.55 ( $\pm 0.13$ )	0.54 ( $\pm 0.14$ )	0.54 ( $\pm 0.14$ )

Table 1:  $\ell_\infty$  loss of the four procedures.

	Best	Cross-Validation	$AV_\infty$	$AV_\infty^m$
$p = 200, u = 0.3$	0.37 ( $\pm 0.09$ )	0.42 ( $\pm 0.10$ )	0.43 ( $\pm 0.11$ )	0.40 ( $\pm 0.09$ )
$p = 500, u = 0.3$	0.44 ( $\pm 0.11$ )	0.50 ( $\pm 0.12$ )	0.49 ( $\pm 0.12$ )	0.48 ( $\pm 0.12$ )
$p = 2000, u = 0.3$	0.64 ( $\pm 0.17$ )	0.67 ( $\pm 0.18$ )	0.67 ( $\pm 0.17$ )	0.66 ( $\pm 0.18$ )

Table 2:  $\ell_\infty$  loss of the four procedures.

**Stability:**  $AV_\infty$  is stable with respect to changes of the constant  $a_c$ : in the above simulations, the sup-norm performances of  $AV_\infty$  with different constants  $a_c \in [0.5, 1]$  differ less than 0.01. Similarly,  $AV_\infty^m$  is stable with respect to the tuning parameter of the tuning parameter.

**Variable selection:** In Table 3, we compare the variable selection performance of Cross-Validation and  $AV_\infty$ . In contrast to Cross-Validation,  $AV_\infty$  allows for a safe threshold of size  $3a_c\hat{\lambda}$  applied to each component (see Remark 2). We therefore report the results of  $AV_\infty$  and of  $AV_\infty^m$  with an additional threshold of size  $3a_c\hat{\lambda}$  applied to each component. We observe that a threshold applied to  $AV_\infty$  can remove a large false positives without increasing the number of false negatives.

	Cross-Validation	$AV_\infty$	$AV_\infty$ with threshold
$p = 200, u = 0$	0.0 ( $\pm 0.0$ )/29.8 ( $\pm 14.4$ )	0.0 ( $\pm 0.0$ )/68.2 ( $\pm 6.0$ )	0.0 ( $\pm 0.0$ )/ 9.1 ( $\pm 5.1$ )
$p = 500, u = 0$	0.0 ( $\pm 0.0$ )/41.3 ( $\pm 18.7$ )	0.0 ( $\pm 0.0$ )/88.3 ( $\pm 5.1$ )	0.0 ( $\pm 0.0$ )/22.1 ( $\pm 8.0$ )
$p = 2000, u = 0$	0.0 ( $\pm 0.0$ )/60.2 ( $\pm 23.4$ )	0.0 ( $\pm 0.0$ )/99.0 ( $\pm 5.5$ )	0.0 ( $\pm 0.0$ )/19.3 ( $\pm 7.0$ )

Table 3: False negatives/false positives for three procedures.

**Computational complexity:** 10-fold Cross-Validation requires the computation 10 Lasso paths, while  $AV_\infty$  requires the computation of only one Lasso path.  $AV_\infty$  is therefore about 10 times faster than 10-fold Cross-Validation. In contrast,  $AV_\infty^m$  additionally requires the computation of additional Lasso problems for the constants  $d_1, \dots, d_p$  and, therefore, computationally the most complex schemes among the three schemes.

### 3.2 Riboflavin production in *B. subtilis*

We now consider variable selection for a data set describing the production of riboflavin (vitamin B<sub>2</sub>) in *B. subtilis* (*Bacillus subtilis*) (Bühlmann, Kalisch and Meier 2014). The data set comprises  $n = 71$  samples of the riboflavin production rates and of the corresponding expressions of  $p = 4088$  genes in *B. subtilis*. As suggested by the theory, we apply  $AV_\infty$  with constant  $a_c = 0.75$  (cf. Lemma 1) and then impose the threshold  $3a_c\hat{\lambda}$  (cf. Remark 2). Similarly, we apply  $AV_\infty^m$  with  $d_1, \dots, d_p$  obtained

$AV_\infty$	$AV_\infty^m$	stability selection	B-TREX
YXLD_at -0.405	YXLD_at -0.478	YXLD_at	YXLD_at
YOAB_at -0.420	YOAB_at -0.389	YOAB_at	YOAB_at
YEBC_at -0.146	YEBC_at -0.224	LYSC_at	YXLE_at
ARGF_at -0.313	YHDS_r_at 0.150		
XHLB_at 0.278			

Table 4: Variable selection results for the riboflavin data set. The first column depicts the genes and the corresponding parameter values yielded by  $AV_\infty$ . The second column depicts the genes and the corresponding parameter values yielded by  $AV_\infty^m$ . The third and fourth column depict the genes returned by approaches based on stability selection and TREX.

via a prior Lasso with  $u_j = \sqrt{\log(p)/n}$  (cf. Lemma 3) and then impose the threshold  $9d_j\tilde{\lambda}/4$  for the  $j$ th component.

The resulting genes and the corresponding parameter values are given in the first and second column of Table 4. We see that these results commensurate with the results from previous approaches based on stability selection (Bühlmann et al. 2014) and B-TREX (Lederer and Müller 2014) given in the third and fourth column.

The results for different choices of  $a_c$  are given in Table 5. These results demonstrate that the resulting genes and parameter values are robust against (even large) changes in the constant  $a_c$ .

Additionally, the results for different choices of  $a_c$  are given in Table 5. Table 4 shows that our results commensurate with the results in (Bühlmann et al. 2014; Lederer and Müller 2014). Moreover, Table 4 shows that  $AV_\infty$  is robust against (even large) changes in the constant  $a_c$  both with respect to the selected variables and the corresponding parameter values.

#### 4. CONCLUSIONS

We have introduced  $AV_\infty$  and  $AV_\infty^m$  to calibrate Lasso for minimal sup-norm loss and have demonstrated that these calibration schemes provide, unlike standard calibration schemes, both optimal theoretical guarantees for  $\ell_\infty$  loss and simple and fast implementations. This means that  $AV_\infty$  and  $AV_\infty^m$  are promising competitors of standard calibration schemes such as Cross-Validation.

$AV_\infty$	$a_c = 0.5$	$AV_\infty$	$a_c = 0.75$	$AV_\infty$	$a_c = 1$
YXLD_at	-0.204	YXLD_at	-0.405	YXLD_at	-0.445
YOAB_at	-0.453	YOAB_at	-0.420	YOAB_at	-0.525
YXLE_at	-0.205	YEBC_at	-0.146	YEBC_at	-0.180
ARGF_at	-0.322	ARGF_at	-0.313	ARGF_at	-0.356
XHLB_at	0.305	XHLB_at	0.278		

Table 5: Stability of  $AV_\infty$  on the riboflavin data set. The three columns depict the genes and the corresponding parameter values yielded by  $AV_\infty$  with different constants  $a_c$ .

Besides sup-norm loss,  $AV_\infty$  and  $AV_\infty^m$  can be of interest for variable selection: Since the tuning parameters provided by  $AV_\infty$  and  $AV_\infty^m$  are upper bounded by the oracle tuning parameter, they can be used - in contrast to standard tuning parameters - for a safe threshold to reduced false positives without increasing false negatives.

We focused on the calibration of Lasso for a broad class of linear regression models; however, we expect that similar models and estimators (such as Square-Root Lasso, for example) can be treated along the same lines. On the contrary, a direct transfer to different losses is not possible, reflecting that calibration needs to be tailored to the task under consideration. This is illustrated for prediction loss in a forthcoming paper.

## ACKNOWLEDGMENTS

We thank Sara van de Geer and Sébastien Loustau for the many helpful discussions and insightful comments.

## References

- Belloni, A., Chernozhukov, V., and Wang, L. (2011), “Square-root lasso: pivotal recovery of sparse signals via conic programming,” *Biometrika*, 98(4), 791–806.
- Bühlmann, P., Kalisch, M., and Meier, L. (2014), “High-Dimensional Statistics with a View Toward Applications in Biology,” *Annual Review of Statistics and Its Application*, 1(1), 255–278.
- Bühlmann, P., and van de Geer, S. (2011), *Statistics for high-dimensional data: Methods, theory and applications*, Springer Series in Statistics Springer.

- Bunea, F. (2008), “Honest variable selection in linear and logistic regression models via  $\ell_1$  and  $\ell_1 + \ell_2$  penalization,” *Electron. J. Stat.*, 2, 1153–1194.
- Chichignoud, M., and Lederer, J. (2014), “A robust, adaptive M-estimator for pointwise estimation in heteroscedastic regression,” *Bernoulli*, 20(3), 1560–1599.
- Dalalyan, A., Hebiri, M., and Lederer, J. (2014), “On the Prediction Performance of the Lasso,” *preprint, arXiv:1402.1700*, .
- Fan, J., and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, 96(456), 1348–1360.
- Hebiri, M., and Lederer, J. (2013), “How Correlations Influence Lasso Prediction,” *IEEE Trans. Inform. Theory*, 59(3), 1846–1854.
- Lederer, J., and Müller, C. (2014), “Don’t fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX,” *preprint, arXiv:1404.0541*, .
- Lepski, O. (1990), “A problem of adaptive estimation in Gaussian white noise,” *Teor. Veroyatnost. i Primenen.*, 35(3), 459–470.
- Lepski, O., Mammen, E., and Spokoiny, V. (1997), “Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors,” *Ann. Statist.*, 25(3), 929–947.
- Lounici, K. (2008), “Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators,” *Electron. J. Stat.*, 2, 90–102.
- Städler, N., Bühlmann, P., and van de Geer, S. (2010), “ $\ell_1$ -penalization for mixture regression models,” *Test*, 19(2), 209–256.
- Sun, T., and Zhang, C. (2012), “Scaled sparse linear regression,” *Biometrika*, 99(4), 879–898.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *J. Roy. Statist. Soc. Ser. B*, 58(1), 267–288.

- van de Geer, S. (2014), “Worst possible sub-directions in high-dimensional models,” *preprint, arXiv:1403.7023*, .
- van de Geer, S., and Bühlmann, P. (2009), “On the conditions used to prove oracle results for the Lasso,” *Electron. J. Stat.*, 3, 1360–1392.
- van de Geer, S., and Lederer, J. (2013), “The Lasso, correlated design, and improved oracle inequalities,” *IMS Collections*, 9, 303–316.
- Zhang, C.-H. (2010), “Nearly unbiased variable selection under minimax concave penalty,” *The Annals of Statistics*, pp. 894–942.